The conjugate prior for discrete hierarchical log-linear models

Jinnan Liu*, Hélène Massam†

Abstract

In the Bayesian analysis of contingency table data, the selection of a prior distribution for either the log-linear parameters or the cell probabilities parameter is a major challenge. Though the conjugate prior on cell probabilities has been defined by Dawid and Lauritzen (1993) for decomposable graphical models, it has not been identified for the larger class of graphical models Markov with respect to an arbitrary undirected graph or for the even wider class of hierarchical log-linear models. In this paper, working with the log-linear parameters used by GLIM, we first define the conjugate prior for these parameters and then derive the induced prior for the cell probabilities: this is done for the general class of hierarchical log-linear models. We show that the conjugate prior has all the properties that one expects from a prior: notational simplicity, ability to reflect either no prior knowledge or a priori expert knowledge, a moderate number of hyperparameters and mathematical convenience. It also has the strong hyper Markov property which allows for local updates within prime components for graphical models.

Keywords: hierarchical log-linear models, conjugate prior, prior specification, hyper Markov property.

AMS 2000 Subject classifications. Primary 62H99; Secondary 62E15

*Department of Statistics, York University, Toronto, M3J 1P3, Canada

†Department of Statistics, York University, Toronto, M3J 1P3, Canada. This author thanks NSERC whose generous support has made this work possible.

1 Introduction

We consider data given under the form of a contingency table representing the classification of n individuals according to a finite set V of criteria. Each criterion $\gamma \in V$ is represented by a variable X_γ which take values in a finite set $\mathcal{I}_\gamma$. We observe the values of the variable $X = (X_\gamma, \gamma \in V)$ in $\mathcal{I} = \times_{\gamma \in V} \mathcal{I}_\gamma$ for these n individuals, and we assume that the resulting $|\mathcal{I}|$ cell counts in the contingency table follow a multinomial distribution. We also assume that the cell probabilities are modeled according to a hierarchical log-linear model parametrized by interaction parameters represented by a vector θ. Since the class of discrete graphical models Markov with respect to an arbitrary undirected graph G is an important subclass of the class of hierarchical log-linear models, we will give special attention to that class througout the paper.

In the Bayesian analysis of contingency table data, the selection of a prior distribution for either the log-linear parameters or the cell probabilities parameter is a major challenge (see Clyde and George, 2004). Priors are usually chosen for their conceptual and computational simplicity and for their ability to represent experts prior beliefs. They are also chosen so that they can conveniently be used for the whole class of log-linear models which includes nondecomposable as well as decomposable graphical models. Moreover their parametrization, that is the hyper-parametrization, should be such that hyper-parameters are compatible across models.

As shown in Dawid and Lauritzen (1993), the conjugate prior for decomposable graphical models, called the hyper Dirichlet and defined for marginal cliques and separators cell probabilities has all of these properties and additionally has the strong hyper Markov property. The latter is very desirable since it allows for local updates within cliques thus simplifying the computation of Bayes factors in a model selection process. The hyper Dirichlet has therefore been used in many studies(see for example Madigan and Raftery, 1994 and Dellaportas and Forster, 1999). However, the hyper Dirichlet is only defined for decomposable graphical models and when it is used as a prior, the corresponding posterior probability for a model is only its probability within this restricted class thus making it difficult to compare it to the posterior probability of another model considered within the wider class of hierarchical log-linear models. Moreover, it appears to have many hyperparameters since a set of parameters has to be chosen for the Dirichlet on each clique and each separator of the graph. In fact, all these hyper-parameters are not independent of each other since they have to be hyper-consistent but the apparently large number of parameters adds a level of complexity to their selection.

Consequently much effort has been devoted to the study of alternative priors. For example, King and Brooks (2001), after a discussion on the advantages and disadvantages of the hyper Dirichlet propose a multivariate normal prior for the log-linear parameters for all hierarchical log-linear models. This prior allows for efficient computation, facilitates prior elicitation and induces a log-normal distribution on the cell probabilities with easy to compute prior mean and covariances.

The aim of this paper is to show that the conjugate prior can also be defined for the wider class of hierarchical log-linear models in a simple way and that it has all the desirable properties that one traditionally wants from a prior. Indeed we will show that experts prior beliefs or lack of any prior information can easily be expressed by an appropriate choice of hyperparameters. The chosen prior is consistent with prior beliefs under both parametrization of the model. The conjugate prior is also hyper Markov thus leading to local updates in graphical models, a property that traditional normal priors on log-linear parameters do not have. Also, the number of hyperparameters is moderate, in fact exactly equal to the number of log-linear parameters plus one and the hyperparameters are hyperconsistent across prime components in graphical models and compatible across models.

In §2, we set our notation and give some preliminary results. We work with the parametrization used by GLIM. It is interesting to note that this parametrization expresses the logarithm of cell probabilities $p(i), i \in \mathcal{I}$, which we regard as functions of i, as the sum of functions $\theta_E(i)$ of i which are in orthogonal subspaces of the space $\mathbb{R}^{\mathcal{I}}$ of functions on $\mathcal{I}$. This orthogonal decomposition of $\log p(i)$ will insure that hyperparameters in the prior are compatible across all models. Our parametrization will also lead us to express the distribution of the marginal cell counts in the contingency table, rather than the cells counts, as an exponential family. Using this exponential family form, we derive, in §3,

the expression of the Diaconis and Ylvisaker (1979) conjugate prior for the log-linear parameters. We give a necessary and sufficient condition for this prior to be proper and two methods to obtain hyperparameters that insure that the prior is proper. In §4, we obtain the expression of the induced conjugate prior for the cell probabilities and in §5, we give the details of the properties we mentioned above. Having the expression of the induced prior on cell probabilities allow us to verify that the choice of hyperparameters in one parametrization (log-linear or cell probabilities) expresses the same prior belief in the other parametrization.

2 The log-linear model

2.1 The parametrization

Let V be the set of criteria. Let $X = (X_\gamma, |\ \gamma \in V)$ such that X_γ takes its values (or levels) in the finite set I_γ of dimension $|I_\gamma|$. When a fixed number of individuals are classified according to the $|V|$ criteria, the data is collected in a contingency table with cells indexed by combination of levels for the $|V|$ variables. We adopt the notation of Lauritzen (1996) and denote a cell by

$$i = (i_\gamma,\ \gamma \in V) \in \mathcal{I} = \times_{\gamma \in V} \mathcal{I}_\gamma.$$

The count in cell i is denoted $n(i)$ and the probability of an individual falling in cell i is denoted $p(i)$. For $E \subset V$, cells in the E-marginal table are denoted

$$i_E \in \mathcal{I}_E = \times_{\gamma \in E} \mathcal{I}_\gamma.$$

The marginal counts are denoted $n(i_E)$. For $n = \sum_{i \in \mathcal{I}} n(i),\ (n) = (n(i),\ i \in \mathcal{I})$ follows a multinomial $\mathcal{M}(n, p(i),\ i \in \mathcal{I})$ distribution with probability density function

$$P(n) \propto \prod_{i \in \mathcal{I}} p(i)^{n(i)}\ . \tag{2.1}$$

Let i^* be a fixed but arbitrary cell which for convenience we take to be the cell indexed by the "lowest levels" for each factor and for convenience again, we denote this level by 0. Therefore i^* can be thought to be the cell

$$i^* = (0, 0, \ldots, 0).$$

Consider the following parametrization

$$\theta_E(i) = \sum_{F \subseteq E} (-1)^{|E \setminus F|} \log p(i_F, i^*_{F^c}) \tag{2.2}$$

where by Moebius inversion

$$p(i) = \exp \sum_{E \subseteq V} \theta_E(i)\ . \tag{2.3}$$

We note that $\theta_\emptyset(i) = \log p(i^*),\ i \in \mathcal{I}$ and we will therefore adopt the notation

$$\theta_\emptyset(i) = \theta_\emptyset,\ \ p(i^*) = p_\emptyset = \exp \theta_\emptyset. \tag{2.4}$$

This parametrization has been used in many papers (see for example Dellaportas and Forster, 1999) and can be found in Lauritzen (1996, p.36).

Let us make an important remark here. Since i^* is fixed, the function

$$i \in \mathcal{I} \mapsto \log p(i_E, i^* E^c)$$

belongs to the factor subspace $\mathcal{U}_E$ (as defined in Lauritzen ,1996, Appendix B.2) of the space $\mathbb{R}^{\mathcal{I}}$ of real-valued function on $\mathcal{I}$ that depend only on i_E. Therefore by Proposition B.4 of Lauritzen (1996) and (2.2) above, $\theta_E(i)$ belongs to the interaction subspace $\mathcal{V}_E$ which gives the "'pure"' contribution of

the interaction between variables in E with the interaction between variables in all $F \subset E$ removed. This means that (2.3) or more precisely its equivalent expression

$$\log p(i) = \sum_{E \subseteq V} \theta_E(i)$$

is the unique expansion of $\log p(i)$ into its orthogonal components in $\mathcal{V}_E, E \subseteq V$. This orthogonal decomposition of $\log p(i)$ is the property that will make the hyperparametrization of the conjugate prior on θ, defined below in (2.15), compatible across models since all models will be expressed in the same orthogonal "basis". Let us now emphasize some other properties of the θ parametrization with the following three lemmas.

Lemma 2.1 *For any $(i) \in \mathcal{I}$ and any $E \subseteq V$, $\theta_E(i)$ depends only on i_E, that is*

$$\theta_E(i) = \theta_E(i_E) \ .$$

Since i^* is fixed, the proof of this first lemma is obvious.

Lemma 2.2 *If i is such that for $\gamma \in E, i_\gamma = i^*_\gamma = 0$, then $\theta_E(i_E) = 0$*

Proof: By definition and since $(i_{F\cup\gamma}, i^*_{(F\cup\gamma)^c}) = (i_F, i^*_{F^c})$ we have

$$\begin{aligned} \theta_E(i) &= \sum_{F\subseteq E\setminus\gamma} (-1)^{|E\setminus F|} \log p(i_F, i^*_{F^c}) - \sum_{F\subseteq E\setminus\gamma} (-1)^{|E\setminus F|} \log p(i_{F\cup\gamma}, i^*_{(F\cup\gamma)^c}) \\ &= \sum_{F\subseteq E\setminus\gamma} (-1)^{|E\setminus F|} \log p(i_F, i^*_{F^c}) - \sum_{F\subseteq E\setminus\gamma} (-1)^{|E\setminus F|} \log p(i_F, i^*_{F^c}) = 0 \ . \end{aligned}$$

□

From this lemma, it follows immediately that our parametrization is the GLIM parametrization that sets to 0 the values of the $E-$ interaction log-linear parameters when at least one index in E is at level 0 (see for example Agresti 1990, p.150) . Therefore, for each $E \subseteq V$, there are only $\prod_{\gamma\in E}(|\mathcal{I}_\gamma| - 1)$ parameters. The next lemma is actually the Hammersley-Clifford theorem and its proof can be found for example in Lauritzen (1996, p.36).

Lemma 2.3 *Assuming all cell probabilities are positive, the distribution of $X = (X_\gamma, \gamma \in V)$ is Markov with respect to the undirected graph G if and only if $\theta_E(i_E) = 0$ whenever $E \subseteq V$ is not complete.*

From this lemma, it follows that the multinomial distribution of the cell counts is Markov with respect to a graph G if and only if θ_E is equal to zero when E is not complete, a well-known property that we recall here (see Darroch, Lauritzen and Speed, 1980).
For notational convenience, we now define

$$\mathcal{E} = \{E \subseteq V, \ \ E \neq \emptyset\}.$$

By Lemma 2.2, for any given $j \in \mathcal{I}$, $\theta_E(j_E) = 0$ if there exists at least one $\gamma \in E$ such that $j_\gamma = 0$. We therefore define for any $j \in \mathcal{I}$

$$\mathcal{E}^*_j = \{E \in \mathcal{E} \mid j_\gamma \neq 0, \forall\gamma \in E\} \ . \tag{2.5}$$

Then, by (2.2), (2.4) and Lemma (2.1),

$$p_\emptyset = 1 - \sum_{i\in\mathcal{I}, i\neq i^*} p(i) = 1 - \sum_{i\in\mathcal{I}, i\neq i^*} \exp \sum_{E\subseteq V} \theta_E(i) = 1 - \sum_{i\in\mathcal{I}, i\neq i^*} p_\emptyset \Big(\exp \sum_{E\in\mathcal{E}^*_i} \theta_E(i_E) \Big) \ .$$

which yields

$$p_\emptyset = \frac{1}{1 + \sum_{j \in \mathcal{I}, j \neq i^*} \exp \sum_{E \in \mathcal{E}_j^*} \theta_E(j_E)} \tag{2.6}$$

and

$$p(i) = \frac{\exp \sum_{E \in \mathcal{E}_i^*} \theta_E(i_E)}{1 + \sum_{j \in \mathcal{I}, j \neq i^*} \exp \sum_{E \in \mathcal{E}_j^*} \theta_E(j_E)} \tag{2.7}$$

and thus all cell probabilities are expressed in terms of the free parameters

$$\theta_E(i_E), E \in \mathcal{E}, i \in \mathcal{I}, i \neq i^*.$$

2.2 The multinomial distribution for discrete data

We now want to give the probability density function of the multinomial distribution under the form of an exponential family. This will be done successively for the saturated model i.e. Markov with respect to a complete graph, for models Markov with respect to an undirected graph G and for general hierarchical log-linear models.
From (2.1), for the saturated model, we have

$$\begin{aligned} P(n) &\propto \prod_{i \in \mathcal{I}} (\exp \sum_{E \subseteq V} \theta_E(i))^{n(i)} = \prod_{i \in \mathcal{I}} \exp n(i) \log \exp \sum_{E \subseteq V} \theta_E(i_E) = \exp \sum_{i \in \mathcal{I}} n(i) \sum_{E \subseteq V} \theta_E(i) \\ &= \exp \sum_{i \in \mathcal{I}} \sum_{E \subseteq V} \theta_E(i_E) \sum_{j \in \mathcal{I}, j_E = i_E} n(j) = \exp \sum_{E \subseteq V} \sum_{i_E \in \mathcal{I}_E} \theta_E(i_E) n(i_E) \\ &= \exp\{ \sum_{E \subseteq V, E \neq \emptyset} \sum_{i_E \in \mathcal{I}_E} \theta_E(i_E) n(i_E) + n\theta_\emptyset) \} \end{aligned}$$

Moreover, we know from Lemma 2.2 that, only those $\theta_E(i_E)$ where $i_\gamma \neq 0, \gamma \in E$ are nonzero. Therefore if, for $E \in \mathcal{E}$ we define

$$\mathcal{I}_E^* = \{ i_E = (i_\gamma, \gamma \in E) \in \mathcal{I}_E,\ i_\gamma \neq 0, \gamma \in E \} \tag{2.8}$$

then the probability density function above becomes

$$P(n) \propto \exp\{ \sum_{E \subseteq V, E \neq \emptyset} \sum_{i_E \in \mathcal{I}_E^*} \theta_E(i_E) n(i_E) + n\theta_\emptyset) \} \tag{2.9}$$

We see that, with the parametrization that we have chosen, the marginal counts $n(i_E)$, rather than the cell counts $n(i)$, appear naturally as random variables . Since the Jacobian of

$$n = (n(i), i \in \mathcal{I}) \mapsto y = (n(i_E),\ E \in \mathcal{E},\ i_E \in \mathcal{I}_E^*)$$

is clearly one, the family of distributions for y is the natural exponential family

$$\mathcal{F}_\mu = \{ f(y;\theta)\mu(y) = \frac{\exp\{\sum_{E \in \mathcal{E}} \sum_{i_E \in \mathcal{I}_E^*} \theta_E(i_E) n(i_E)\}}{\left(1 + \sum_{i \in \mathcal{I}, i \neq i^*} \exp \sum_{F \in \mathcal{E}_i^*} \theta_F(i_F)\right)^n} \mu(y),\ \ \theta \in \mathbb{R}^{\prod_{E \in \mathcal{E}} \prod_{\gamma \in E} (|I_\gamma| - 1)} \} , \tag{2.10}$$

where μ is a reference measure of no particular interest to us here. This gives us the density for the saturated model.

When G is an arbitrary undirected graph let

$$\mathcal{D} = \{ E \in \mathcal{E} \mid E \text{ complete} \}$$

and for any given $j \in \mathcal{I}$

$$\mathcal{D}_j^* = \{ D \in \mathcal{D} \mid j_\gamma \neq 0, \forall \gamma \in D \}, \tag{2.11}$$

From Lemma 2.3, to obtain the expression of the cell probabilities and of the family of multinomial distributions Markov with respect to G it suffices to equate $\theta_E(i)$ to 0 for $E \notin \mathcal{D}$ and for all (i). We therefore have

$$p_\emptyset = \frac{1}{1 + \sum_{j \in \mathcal{I}, j \neq i^*} \exp \sum_{D \in \mathcal{D}_j^*} \theta_D(j_D)} \tag{2.12}$$

$$p(i) = \frac{\exp \sum_{D \in \mathcal{D}_i^*} \theta_D(i_D)}{1 + \sum_{j \in \mathcal{I}, j \neq i^*} \exp \sum_{D \in \mathcal{D}_j^*} \theta_D(j_D)} \ , \tag{2.13}$$

where it is important to note that in (2.13) not all $p(i)$ are free parameters since

$$\text{for} \quad E \notin \mathcal{D}, \quad \theta_E(i_E) = 0$$

which implies that for $E \notin \mathcal{D}$, $p(i_E, i^*_{E^c})$ is function of $p(i_F, i^*_{F^c}), F \subset E, \ F \in \mathcal{D}$. Only cell probabilities of the form $p(i_D, i^*_{D^c}), \ i_D \in \mathcal{I}^*_D, \ D \in \mathcal{D}$ will be free probabilities and form the cell probability parameter

$$p = (p(i_D, i^*_{D^c}), \ D \in \mathcal{D}, \ i_D \in \mathcal{I}^*_D,) \ \text{ with } \ p(i_D, i^*_{D^c}) \ \text{ as in } \ (2.3) \ . \tag{2.14}$$

of the multinomial distribution Markov with respect to G for graphical models, or of the hierarchical log-linear model. The corresponding log-linear parameters are obviously

$$\theta = (\theta_D(i_D), \ D \in \mathcal{D}, \ i_D \in \mathcal{I}^*_D) \ \text{ with } \ \theta_D(i_D) \ \text{ as in } \ (2.2) \ . \tag{2.15}$$

Moreover, the family of multinomial distribution Markov with respect to G for

$$y = (n(i_D), \ D \in \mathcal{D}, \ i_D \in \mathcal{I}^*_D)$$

is

$$\mathcal{F}_{\mu_G} \quad = \quad \{f_G(y;\theta)\mu_G(y) = \frac{\exp\{\sum_{D \in \mathcal{D}} \sum_{i_D \in \mathcal{I}^*_D} \theta_D(i_D) n(i_D)\}}{\Big(1 + \sum_{i \in \mathcal{I}, i \neq i^*} \exp \sum_{D \in \mathcal{D}_i^*} \theta_D(i_D)\Big)^n} \mu_G(y), \\ \theta \in \mathbb{R}^{\prod_{D \in \mathcal{D}} \prod_{\gamma \in D} (|\mathcal{I}_\gamma| - 1)}\} \tag{2.16}$$

where $\theta = (\theta_D(i_D), D \in \mathcal{D}, i_D \in \mathcal{I}_D)$ and μ_G is a reference measure of no particular interest to us here. Densities in $\mathcal{F}_{\mu_G}$ will be written under the natural exponential family form

$$f_G(y;\theta) \quad = \quad \exp\{\sum_{D \in \mathcal{D}} \sum_{i_D \in \mathcal{I}^*_D} \theta_D(i_D) n(i_D)\} - n \log\Big(1 + \sum_{i \in \mathcal{I}, i \neq i^*} \exp \sum_{D \in \mathcal{D}_i^*} \theta_D(i_D)\Big) \tag{2.17}$$

When the model is a hierarchical log-linear model, let $\mathcal{D}$ be the set of subsets of V representing the set of all possible interactions in the given model, which we will call the generating set. Then, the expression of the cell probabilities and of the multinomial distribution for this model is the same as in (2.12), (2.13) and (2.16) but with $\mathcal{D}$ representing the generating set for the model.

2.3 The multinomial distribution for binary data

We consider here the important special case of binary data, because it occurs often in practice and also because in this case, the notation is somewhat simpler. When the variables $X_\gamma, \gamma \in V$ can only take two values 0 or 1, there is only one cell i in each $\mathcal{I}^*_E, E \in \mathcal{E}$ and therefore each cell $i = (i_\gamma, \gamma \in V)$ can be indexed by $E = \{\gamma \in V : i_\gamma = 1\}$ for $E \in \mathcal{E} \cup \emptyset$ such that $i_\gamma \neq 0, \gamma \in E$. The correspondence between $\mathcal{I}$ and $\mathcal{E} \cup \emptyset$ is one to one. For $i = (i_\gamma = 1, \gamma \in E, i_\gamma = 0, i \notin E)$, we will therefore use the notation

$$p_E = p(i) \ \text{ for and } \ \theta_E = \theta_E(i_E).$$

The relation (2.2) becomes

$$\theta_F = \sum_{E \subseteq F} (-1)^{|F \setminus E|} \log p_E = \log \prod_{E \subseteq F} p_E^{(-1)^{|F \setminus E|}}, \quad F \in \mathcal{E} \cup \emptyset \tag{2.18}$$

Or equivalently by Moebius inversion

$$\log p_F = \sum_{E \subseteq F} \theta_E, \quad F \in \mathcal{E} \quad \text{with} \quad \log p_\emptyset = \theta_\emptyset \ , \tag{2.19}$$

Then (2.6), (2.7) and (2.16) become respectively

$$p_\emptyset = \frac{1}{\left(1 + \sum_{E \in \mathcal{E}} \exp\{\sum_{F \subseteq E, F \in \mathcal{D}} \theta_F\}\right)} \tag{2.20}$$

$$p_E = \frac{\exp \sum_{F \subseteq E, F \in \mathcal{D}} \theta_F}{\left(1 + \sum_{H \in \mathcal{E}} \exp\{\sum_{F \subseteq H, F \in \mathcal{D}} \theta_F\}\right)}, \quad E \in \mathcal{E} \tag{2.21}$$

and

$$\mathcal{F}_{\mu_G} = \{f(y;\theta,G)\mu_G(y) = \exp\Big(\sum_{D \in \mathcal{D}} \theta_D y_D - n \log(1 + \sum_{E \in \mathcal{E}} \exp(\sum_{D \subseteq E, D \in \mathcal{D}} \theta_D))\Big)\, \mu_G(y) \tag{2.22}$$
$$\theta = (\theta_D, D \in \mathcal{D}) \in \Theta_G = \mathbb{R}^{|\mathcal{D}|}\}$$

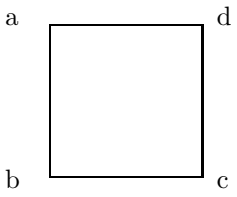
where $\mathcal{D}$ is equal to $\mathcal{E}$, the set of complete subsets of V in G or the generating set for the hierarchical model for, respectively, the saturated model, graphical model with respect to G or the hierarchical model.
We note that (2.14) and (2.15) become

$$p = (p_D, D \in \mathcal{D}) \ \text{ and } \ \theta = (\theta_D, D \in \mathcal{D})$$

respectively.

2.4 An example

We consider the case where $X = (X_a, X_b, X_c, X_d)$ is Markov with respect to the four-cycle as given below and where the variables are binary.



We then have

$$\begin{aligned} \mathcal{D} &= \{a, b, c, d, ab, bc, cd, da\} \\ \mathcal{E} &= \{a, b, c, d, ab, bc, cd, da, ac, bd, abc, bcd, cda, dab, abcd\} \end{aligned}$$

The linear constraints on $\theta_E, E \notin \mathcal{D}$ are

$$\theta_{ac} = \theta_{bd} = \theta_{abc} = \theta_{bcd} = \theta_{cda} = \theta_{dab} = \theta_{abcd} = 0$$

and using (2.18), we obtain $p_E, E \notin \mathcal{D}$ in terms of $p = (p_E,\ E \in \mathcal{D})$ as follows

$$p_{ac} = \frac{p_a p_c}{p_\emptyset},\ p_{bd} = \frac{p_b p_d}{p_\emptyset},\ p_{abc} = \frac{p_{ab} p_{bc}}{p_b},\ p_{bcd} = \frac{p_{bc} p_{cd}}{p_c},\ p_{cda} = \frac{p_{cd} p_{da}}{p_d},\ p_{dab} = \frac{p_{da} p_{ab}}{p_a}\ ,$$
$$p_{abcd} = \frac{p_{ab} p_{bc} p_{cd} p_{da} p_\emptyset}{p_a p_b p_c p_d}\ .$$

The cell probability parameters of the multinomial distribution Markov with respect to the four-cycle above can be written in terms of θ as

$$\begin{aligned}
p_\emptyset^{-1} &= 1 + e^{\theta_a} + e^{\theta_b} + e^{\theta_c} + e^{\theta_d} + e^{\theta_a+\theta_b+\theta_{ab}} + e^{\theta_b+\theta_c+\theta_{ab}} + e^{\theta_c+\theta_d+\theta_{cd}} + e^{\theta_d+\theta_a+\theta_{da}} \\
&\quad + e^{\theta_a+\theta_b+\theta_c+\theta_{ab}+\theta_{bc}} + e^{\theta_b+\theta_c+\theta_d+\theta_{bc}+\theta_{cd}} + e^{\theta_c+\theta_d+\theta_a+\theta_{cd}+\theta_{da}} + e^{\theta_d+\theta_a+\theta_b+\theta_{da}+\theta_{ab}} \\
&\quad + e^{\theta_a+\theta_b+\theta_c+\theta_d+\theta_{ab}+\theta_{bc}+\theta_{cd}+\theta_{da}} \\
p_D &= p_\emptyset e^{\theta_D},\ D \in \{a, b, c, d, \}\ \text{ with} \\
p_{ab} &= p_\emptyset e^{\theta_a+\theta_b+\theta_{ab}},\ p_{bc} = p_\emptyset e^{\theta_b+\theta_c+\theta_{bc}},\ p_{cd} = p_\emptyset e^{\theta_c+\theta_d+\theta_{cd}},\ p_{da} = p_\emptyset e^{\theta_d+\theta_a+\theta_{da}}, \\
p_{abc} &= \frac{p_{ab} p_{bc}}{p_b},\ p_{bcd} = \frac{p_{bc} p_{cd}}{p_c},\ p_{cda} = \frac{p_{cd} p_{da}}{p_d},\ p_{dab} = \frac{p_{da} p_{ab}}{p_a},\ p_{abcd} = p_\emptyset \frac{p_{ab} p_{bc} p_{cd} p_{da}}{p_a p_b p_c p_d}
\end{aligned}$$

3 The conjugate prior for the log-linear parameter θ

From (2.17), it is clear that, for the three nested classes of models considered in this paper, graphical with respect to G decomposable, graphical with respect to an arbitrary undirected G and hierarchical, the probability density function for the marginal counts y can be written under an exponential family form and therefore the form of the conjugate prior for θ is given immediately (see Diaconis and Ylvisaker, 1979) by

$$\pi_G(\theta|s,\alpha) = I_G(s,\alpha)^{-1} \exp\{\sum_{D\in\mathcal{D}} \sum_{i_D \in \mathcal{I}_D^*} \theta_D(i_D) s(i_D)\} - \alpha \log\Big(1 + \sum_{i\in\mathcal{I}, i\neq i^*} \exp \sum_{D\in\mathcal{D}_i^*} \theta_D(i_D)\Big) \quad (3.1)$$

where $I_G(s,\alpha)$ is the normalising constant

$$I_G(s,\alpha) = \int_{\mathbb{R}} \prod_{D\in\mathcal{D}} \prod_{\gamma\in D^{(|I_\gamma|-1)}} \pi_G(\theta|s,\alpha)\ d\theta \quad (3.2)$$

and where, as usual, $\mathcal{D}$ is equal to $\mathcal{E}$ when the model is saturated, to the set of complete subsets of G when the model is graphical Markov with respect to G and to the generating set for the model when the model is hierarchical.
In order to be able to use this prior in practice, we need to answer a number of questions. The first basic question is to know for which values of the hyper parameters (s,α) where

$$s = (s(i_D),\ D \in \mathcal{D}, i_D \in \mathcal{I}_D^*)\ \in\ \mathbb{R}^{\prod_{D\in\mathcal{D}} \prod_{\gamma\in D^{(|I_\gamma|-1)}}} \quad \text{and} \quad \alpha \in \mathbb{R}$$

the distribution is proper, i.e. when does $I_G(s,\alpha) < +\infty$ hold. We will now give a necessary and sufficient condition for (3.1) to be proper as well as two practical methods to construct hyper parameters (s,α) such that it is proper. The next set of questions is concerned with the properties of this prior distribution in practice such as ease of prior specification, hyper Markov property. These questions will be addressed in §5.

3.1 A necessary and sufficient condition for the prior to be proper

Lemma 3.1 *The prior distribution (3.1) is properif and only if* (s,α) *belongs to*

$$\Pi_G = \{(s,\alpha) \mid \alpha > 0, (\frac{s(i_D)}{\alpha} = \sum_{j\in\mathcal{I}, j_D = i_D} p(j),\ D \in \mathcal{D},\ i_D \in \mathcal{I}_D^*))\ \textit{with}\ p(j)\ \textit{as in (2.21)}\} \ . \quad (3.3)$$

Proof: Since the parameter space of (2.16) is $\Theta_G = \mathbb{R}^{\prod_{D\in\mathcal{D}}\prod_{\gamma\in D}(|\mathcal{I}_\gamma|-1)}$, by Theorem 1 of Diaconis and Ylvisaker (1974), a necessary and sufficient condition for (3.2) to be finite is that $\alpha > 0$ and $\frac{n}{\alpha}s = \frac{n}{\alpha}(s(i_D),\ D \in \mathcal{D},\ i_D \in \mathcal{I}_D)$ is in the interior of the convex hull of the support of μ_G. Since the Laplace transform

$$L_{\mu_G}(\theta) = \Big(1 + \sum_{i\in\mathcal{I}, i\neq i^*} \exp \sum_{D\in\mathcal{D}_i^*} \theta_D(i_D)\Big)^n$$

is defined for Θ_G which is an open set, the interior of the convex hull of the support of μ_G is equal to the mean space M_G of $\mathcal{F}_{\mu_G}$. We therefore want to identify M_G. Let $k_{\mu_G}(\theta) = \log L_{\mu_G}(\theta)$. Since $\mathcal{F}_{\mu_G}$ is a natural exponential family with parameter $\theta \in \Theta_G$, we have

$$M_G = \{m = (m(i_D),\ D \in \mathcal{D},\ i_D \in \mathcal{I}_D^*) \mid m(i_D) \quad = \quad E(n(i_D)) = n\frac{dk_{\mu_G}(\theta)}{d\theta(i_D)} = n\sum_{j\in\mathcal{I}, j_D = i_D} p(j)\ \} \quad (3.4)$$

where $p(j)$ is as in (2.21). It follows immediately that $(s,\alpha) \in \Pi_G$ is a necessary and sufficient condition for $\pi_G(\theta|s,\alpha)$ to be proper. □

From the lemma above, it is clear that in order to belong to Π_G, (s,α) must satisfy

$$\alpha \geq \max_{D\in\mathcal{D}} s_D \quad \text{and} \quad s_D > s_E \ \text{ for } \ D \subset E,\ D, E \in \mathcal{D}\ .$$

However, this condition is not sufficient since (s,α) must also be such that the $p(j)$ in (3.3) satisfy the conditions $\theta_E(j_E) = 0,\ E \notin \mathcal{D}$.

3.2 Two methods to construct $(s,\alpha) \in \Pi_G$

From (3.3), we immediately obtain the following method to construct hyper parameters (s,α) which are in Π_G:

1. Choose an arbitrary $\theta = (\theta(i_D),\ D \in \mathcal{D}|\ i_D \in \mathcal{I}_D^*)$
2. Compute $p(i)$ according to (2.13).
3. Compute $\frac{s(i_D)}{\alpha} = \sum_{j\in\mathcal{I}, j_D = i_D} p(j)$ for $D \in \mathcal{D},\ i_D \in \mathcal{I}_D^*$.
4. Take $\alpha = 1$.

Another practical way to construct $(s,\alpha) \in \Pi_G$ is to start with a " prior contingency table" with all cell counts $n(i)$ positive. With n denoting the total count in the given contingency table, the maximum likelihood estimate $\hat{p}$ of p satisfying the equations

$$n(i_D) = n\sum_{j\in\mathcal{I}, j_D = i_D} \hat{p}(j),\ D \in \mathcal{D},\ i_D \in \mathcal{I}_D^*$$

and the constraints of the model, exists and therefore we can take

$$\alpha = n,\ \ s(i_D) = n(i_D),\ D \in \mathcal{D},\ i_D \in \mathcal{I}_D^*$$

thus obtaining hyperparameters in Π_G.

We note here that these hyperparameters are consistent across models since the "marginal counts" do not change when we take different models. Marginal counts do not change either when we take marginal or conditional models.

4 The induced prior on the cell probabilities

In this section, we will give the expression of the induced conjugate prior in terms of p, the cell probability parameter, first for graphical models Markov with respect to a decomposable G thus making the link between the hyper Dirichlet and our conjugate prior, then for models Markov with respect to an arbitrary graph G and finally for general hierarchical models.

4.1 The conjugate prior when G is decomposable

For G decomposable with set of cliques $\mathcal{C} = \{C_l, l = 1, \ldots, k\}$ and set of minimal separators $\mathcal{S} = \{S_l, i = 2, \ldots, k\}$, Dawid and Lauritzen (1993) defined the conjugate prior in terms of cell probabilities and called it the hyper Dirichlet distribution. Its density is expressed in terms of

$$p^{C_l}(i_D, i^*_{D^c}),\ D \subseteq C_l, l = 1, \ldots, k, \quad p^{S_l}(i_D, i^*_{D^c}),\ D \subseteq S_l,\ l = 2, \ldots, k, \quad D \in \mathcal{D}\ , i_D \in \mathcal{I}^*_D\ , \quad (4.1)$$

the cell probabilities for the cliques and separators marginal tables, respectively. Note that in this subsection, for $D \subseteq C_l$ or $D \subseteq S_l, D^c$ denotes the complement of D in C_l or S_l respectively. The density of the hyper Dirichlet is equal to

$$\frac{\prod_{l=1}^k \mathrm{Dir}_{C_l}(p^{C_l}_\emptyset, p^{C_l}(i_D, i^*_{D^c}); \alpha^{C_l}_\emptyset, \alpha^{C_l}(i_D, i^*_{D^c}), D \in \mathcal{D}^{C_l}, i_D \in \mathcal{I}^*_D)}{\prod_{l=2}^k \mathrm{Dir}_{S_l}(p^{S_l}_\emptyset, p^{S_l}(i_D, i^*_{D^c}); \alpha^{S_l}(i_D, i^*_{D^c}), D \in \mathcal{D}^{S_l}, i_D \in \mathcal{I}^*_D)} \quad (4.2)$$

with

$$\begin{aligned} & \mathrm{Dir}_{C_l}(p^{C_l}_\emptyset, p^{C_l}(i_D, i^*_{D^c}); \alpha^{C_l}(i_D, i^*_{D^c}), D \in \mathcal{D}^{C_l}, i_D \in \mathcal{I}^*_D) \\ = & \frac{\Gamma(\alpha^{C_l}_\emptyset + \sum_{D \in \mathcal{D}^{C_l}} \sum_{i_D \in \mathcal{I}^*_D} \alpha^{C_l}(i_D, i^*_{D^c}))}{\Gamma(\alpha^{C_l}_\emptyset) \prod_{D \in \mathcal{D}^{C_l}, i_D \in \mathcal{I}^*_D} \Gamma(\alpha^{C_l}(i_D, i^*_{D^c}))} (p^{C_l}_\emptyset)^{\alpha^{C_l}_\emptyset - 1} \prod_{D \in \mathcal{D}^{C_l}, i_D \in \mathcal{I}^*_D} (p^{C_l}(i_D, i^*_{D^c}))^{\alpha^{C_l}(i_D, i^*_{D^c}) - 1} \end{aligned}$$

with a similar expression for Dir_{S_l} and where the hyper parameters

$$(\alpha^{C_l}(i_D, i^*_{D^c}), D \in \mathcal{D}^{C_l}, i_D \in \mathcal{I}^*_D) \text{ and } (\alpha^{S_l}(i_D, i^*_{D^c}), D \in \mathcal{D}^{S_l}, i_D \in \mathcal{I}^*_D) \quad (4.3)$$

are hyperconsistent.

Since $\pi_G(\theta|s, \alpha)$ in (3.1) is the conjugate prior to the multinomial Markov with respect to G, it must coincide with the hyper Dirichlet when G is decomposable. The aim of this subsection is to give the correspondence between the parameter (s, α) and the parameters of the hyper Dirichlet explicitly.

The probabilities in (4.1) are not all free variables since, by the Markov properties of the multinomial distribution,

$$p(i) = \frac{\prod_{l=1}^k p^{C_l}(i_{C_l})}{\prod_{l=2}^k p^{S_l}(i_{S_l})}$$

and therefore some are functions of the others. Let $\mathcal{D}^{C_i}$ and $\mathcal{D}^{S_i}$ denote the set of nonempty subsets of C_i and S_i respectively. We can choose the free marginal probabilities to be

$$p^G = (p^{C_l}(i_D, i^*_{D^c}), D \in \mathcal{D}^{C_l} \setminus \cup_{j=2}^k \mathcal{D}^{S_j}, l = 1, \ldots, k, p^{S_l}(i_D, i^*_{D^c}), D \in \mathcal{D}^{S_l}, l = 2, \ldots, k, i_D \in \mathcal{I}^*_D)\ . \quad (4.4)$$

The Jacobian of the change of variable $\theta \mapsto p^G$ is given in the following lemma.

Lemma 4.1 *The Jacobian of the change of variables from $\theta = (\theta(i_D), D \in \mathcal{D}, i_D \in \mathcal{I}^*_D)$ as given in (2.2) to p^G as given in (4.4) is*

$$\Big|\frac{d\theta}{dp^G}\Big|^{-1} = \frac{\prod_{l=1}^k p^{C_l}_\emptyset \prod_{D \in \mathcal{D}^{C_l}} \prod_{i_D \in \mathcal{I}^*_D} p^{C_l}(i_D)}{\prod_{l=2}^k p^{S_l}_\emptyset \prod_{D \in \mathcal{D}^{S_l}} \prod_{i_D \in \mathcal{I}^*_D} p^{S_l}(i_D)} \quad (4.5)$$

The proof of this lemma is given in the Appendix. The correspondence between (s, α) and (4.3) is given by the following proposition.

Proposition 4.1 *When the graph G is decomposable with set of cliques $(C_i, i = 1, \ldots, k)$ and sets of minimal separators $(S_i, i = 2, \ldots, k)$, the conjugate prior induced from (3.1) is identical to the hyper Dirichlet (4.2) with hyper parameters (4.3) where*

$$\alpha^{C_l}(i_D, i^*_{D^c}) = \sum_{C_l \supseteq F \supseteq D} \sum_{j_F \in \mathcal{I}^*_F \mid (j_F)_D = i_D} (-1)^{|F \setminus D|} s(j_F) \qquad \alpha^{C_i}_\emptyset = \alpha + \sum_{D \subseteq C_i} (-1)^{|D|} \sum_{i \in \mathcal{I}^*_D} s(i_D) \quad (4.6)$$

$$\alpha^{S_l}(i_D, i^*_{D^c}) = \sum_{S_l \supseteq F \supseteq D} \sum_{j_F \in \mathcal{I}^*_F \mid (j_F)_D = i_D} (-1)^{|F \setminus D|} s(j_F) \qquad \alpha^{S_l}_\emptyset = \alpha + \sum_{D \subseteq S_l} (-1)^{|D|} \sum_{i \in \mathcal{I}^*_D} s(i_D) \; . \quad (4.7)$$

Moreover

$$I_G(s, \alpha) = \frac{\prod_{l=1}^k \Gamma(\alpha^{C_l}_\emptyset) \prod_{D \in \mathcal{D}^{C_l}} \prod_{i_D \in \mathcal{I}^*_D} \Gamma(\alpha^{C_l}(i_D, i^*_{D^c}))}{\prod_{l=2}^k \Gamma(\alpha^{S_l}_\emptyset) \prod_{D \in \mathcal{D}^{S_l}} \prod_{i_D \in \mathcal{I}^*_D} \Gamma(\alpha^{S_l}(i_D, i^*_{D^c}))} \quad (4.8)$$

Proof: Since the distribution of Y in (2.22) is Markov with respect to G, we have that

$$p(i) = \frac{\prod_{l=1}^k p^{C_l}(i_{C_l})}{\prod_{l=2}^k p^{S_l}(i_{S_l})} \; . \quad (4.9)$$

Then

$$\begin{aligned}
\theta_E(i_E) &= \sum_{F \subseteq E} (-1)^{|E \setminus F|} \log p(i_F, i^*_{F^c}) \\
&= \sum_{F \subseteq E} (-1)^{|E \setminus F|} \Big(\sum_{l=1}^k \log p^{C_l}(i_{F \cap C_l}, i^*_{F^c \cap C_l}) - \sum_{l=2}^k \log p^{S_l}(i_{F \cap S_l}, i^*_{F^c \cap S_l}) \Big) \\
&= \sum_{l=1}^k \Big(\sum_{F \subseteq E} (-1)^{|E \setminus F|} \log p^{C_l}(i_{F \cap C_l}, i^*_{F^c \cap C_l}) \Big) - \sum_{l=2}^k \Big(\sum_{F \subseteq E} (-1)^{|E \setminus F|} \log p^{S_l}(i_{F \cap S_l}, i^*_{F^c \cap S_l}) \Big) \\
&= \sum_{l=1}^k \theta^{C_l}_E(i_{E \cap C_l}) - \sum_{l=2}^k \theta^{S_l}_E(i_{E \cap S_l})
\end{aligned}$$

If $E \subseteq C_l$, $E \cap C_l = E$ and $\theta^{C_l}_E(i_{E \cap C_l}) = \theta^{C_l}_E$. If $E \not\subseteq C_l$, then by Lemma 2.2, $\theta^{C_l}_E(i_{E \cap C_l}) = 0$ and similarly for $\theta^{S_l}_E(i_{E \cap S_l})$. We therefore have

$$\theta_D(i_D) = \sum_{l=1}^k \theta^{C_i}_D(i_D) - \sum_{l=2}^k \theta^{S_l}_D(i_D), \; D \in \mathcal{D} \quad (4.10)$$

where

$$\begin{aligned}
\theta^{C_i}_D(i_D) &= \sum_{F \subseteq D, F \in \mathcal{D}^{C_i}_0} (-1)^{|D \setminus F|} \log p^{C_i}(i_F, i^*_{F^c}), \quad \text{for } \; D \subseteq C_i \\
\theta^{C_i}_\emptyset &= \log p^{C_i}_\emptyset \\
\theta^{C_i}_D(i_D) &= 0 \quad \text{for } \; D \not\subseteq C_i
\end{aligned}$$

and similar expressions for $\theta^{S_i}(i_D)$ (see also Consonni and Leucari, 2005 for the derivation of these formulas in the case of bivariate data). From (4.9), we also have

$$\log \Big(1 + \sum_{j \in \mathcal{I}, j \neq i^*} \exp \sum_{D \in \mathcal{D}^*_j} \theta_D(j_D) \Big) = -\log p_\emptyset = -\Big(\sum_{l=1}^k \log p^{C_l}_\emptyset - \sum_{l=2}^k \log p^{S_l}_\emptyset \Big) \quad (4.11)$$

Therefore (3.1) can be written as

$$
\begin{aligned}
\pi_G(\theta(p)|s,\alpha) &\propto \frac{\prod_{l=1}^{k}\exp\Big\{\sum_{D\in\mathcal{D}^{C_l}}\sum_{i_D\in\mathcal{I}_D^*}\Big(\sum_{E\subseteq D}(-1)^{|D\setminus E|}\log p^{C_l}(i_E,i^*_{E^c})\Big)s(i_D)+\alpha\log p_\emptyset^{C_l}\Big\}}{\prod_{l=2}^{k}\exp\{\sum_{D\in\mathcal{D}^{S_l}}\sum_{i_D\in\mathcal{I}_D^*}\Big(\sum_{E\subseteq D}(-1)^{|D\setminus E|}\log p^{S_l}(i_E,i^*_{E^c})\Big)s(i_D)+\alpha\log p_\emptyset^{S_l}\}} \\
&= \frac{\prod_{l=1}^{k}\exp\Big\{\sum_{E\in\mathcal{D}^{C_l}}\sum_{i_E\in\mathcal{I}_E^*}\alpha^{C_l}(i_E,i^*_{E^c})\log p^{C_l}(i_E,i^*_{E^c})+\alpha_\emptyset^{C_l}\log p_\emptyset^{C_l}\Big\}}{\prod_{l=2}^{k}\exp\Big\{\sum_{E\in\mathcal{D}^{S_l}}\sum_{i_E\in\mathcal{I}_E^*}\alpha^{S_l}(i_E,i^*_{E^c})\log p^{S_l}(i_E,i^*_{E^c})+\alpha_\emptyset^{S_l}\log p_\emptyset^{S_l}\Big\}} \\
&= \frac{\prod_{l=1}^{k}(p_\emptyset^{C_l})^{\alpha_\emptyset^{C_l}}\prod_{E\in\mathcal{D}^{C_l}}\prod_{i_E\in\mathcal{I}_E^*}(p^{C_l}(i_E,i^*_{E^c}))^{\alpha^{C_l}(i_E,i^*_{E^c})}}{\prod_{l=2}^{k}(p_\emptyset^{S_l})^{\alpha_\emptyset^{S_l}}\prod_{E\in\mathcal{D}^{S_l}}\prod_{i_E\in\mathcal{I}_E^*}(p^{S_l}(i_E,i^*_{E^c}))^{\alpha^{S_l}(i_E,i^*_{E^c})}}
\end{aligned}
\tag{4.12}
$$

where $\alpha^{C_l}(i_E,i^*_{E^c})$, $\alpha^{S_l}(i_E,i^*_{E^c})$, $\alpha_\emptyset^{C_l}$ and $\alpha_\emptyset^{S_l}$ are as defined in (4.6) and (4.7).
The induced prior on p is obtained by multiplying (4.12) by the Jacobian (8.2) and it follows immediately that it is the hyper Dirichlet with hyper parameters as given in (4.6) and (4.7).

The expression of (4.8) is obtained by noticing that for any C_i or S_i,

$$
\alpha_\emptyset^{C_l}+\sum_{E\in\mathcal{D}^{C_l}}\sum_{i_E\in\mathcal{I}_E^*}\alpha^{C_i}(i_E,i^*_{E^c})=\alpha=\alpha_\emptyset^{S_l}+\sum_{E\in\mathcal{E}^{S_i}}\sum_{i_E\in\mathcal{I}_E^*}\alpha^{S_l}(i_E,i^*_{E^c})\ .
$$

This completes the proof. □

4.2 The conjugate prior when G is arbitrary

To obtain the conjugate prior in terms of p, we need to compute the Jacobian $\frac{d\theta}{dp}$ of the transformation from θ to p as defined respectively in (2.15) and (2.14). Before doing so, we need to define the following quantities. For $C\in\mathcal{D},\ H\in\mathcal{E}$, let

$$
F(i_C,j_H)=\begin{cases}(-1)^{|C|-1} & \text{if } (j_H)_C=i_C\\ 0 & \text{otherwise}\end{cases}\qquad \text{for } i_C\in\mathcal{I}_C^*,\ j_H\in\mathcal{I}_H^*\ . \tag{4.13}
$$

These $F(i_C,j_H)$ can be gathered in a $\prod_{D\in\mathcal{D}}|\mathcal{I}_D^*|\times\prod_{H\in\mathcal{E}}|\mathcal{I}_H^*|$ matrix F where the rows are indexed by $i_D\in\mathcal{I}_D^*, D\in\mathcal{D}$ and the columns by $j_H\in\mathcal{I}_H^*, H\in\mathcal{E}\cup\{\emptyset\}$.

For example, in the case of binary data for $\mathcal{D}$ and $\mathcal{E}$ as given in §2.4 the matrix F is

$$
F=\begin{pmatrix}
0&1&0&0&0&1&0&0&1&1&0&1&0&1&1&1\\
0&0&1&0&0&1&1&0&0&0&1&1&1&0&1&1\\
0&0&0&1&0&0&1&1&0&1&0&1&1&1&0&1\\
0&0&0&0&1&0&0&1&1&0&1&0&1&1&1&1\\
0&0&0&0&0&-1&0&0&0&0&0&-1&0&0&-1&-1\\
0&0&0&0&0&0&-1&0&0&0&0&-1&-1&0&0&-1\\
0&0&0&0&0&0&0&-1&0&0&0&0&-1&-1&0&-1\\
0&0&0&0&0&0&0&0&-1&0&0&0&0&-1&-1&-1
\end{pmatrix}
\tag{4.14}
$$

We also need the following two lemmas. Their proof is given in the Appendix.

Lemma 4.2 *Let G be a nondecomposable prime graph. For the matrix F as described in (4.13), the sum of the entries in each column $j_H,\ j\in\mathcal{I}_H^*,\ H\in\mathcal{E}$ is such that*

$$
\sum_{i_C\in\mathcal{I}_C^*,C\in\mathcal{D}}F(i_C,j_H)=1 \tag{4.15}
$$

if and only if H, as an induced subgraph of G, is decomposable and connected.

We are now in a position to give the expression of the Jacobian. Let

$$\mathcal{U} = \{F \in \mathcal{E} \,|F \text{ is either nondecomposable or nonconnected }\}$$

and we also write $\mathcal{U}_0 = \mathcal{U} \cup \{\emptyset\}$.

Lemma 4.3 *Let*

$$a(j_H) = \Big(\sum_{i_C \in \mathcal{I}^*_C, C \in \mathcal{D}} F(i_C, j_H) - 1 \Big), \;\; j \in \mathcal{I}^*_H, \; H \in \mathcal{E} \cup \emptyset \; . \tag{4.16}$$

The Jacobian of the transformation

$$p = (p(i_D, i^*_{D^c}), \; i_D \in \mathcal{I}^*_D, \; D \in \mathcal{D}) \mapsto \; \theta = (\theta_D(i_D), \; i_D \in \mathcal{I}^*_D, \; D \in \mathcal{D}), \tag{4.17}$$

where p is as given in (2.14) and θ as in (2.2) is

$$\begin{aligned} J = \Big|\frac{dp}{d\theta}\Big| &= \Big(\prod_{D \in \mathcal{D}} \prod_{i_D \in \mathcal{I}^*_D} p(i_D, i^*_{D^c}) \Big) \Big(- \sum_{H \in \mathcal{U}_0} \sum_{j_H \in \mathcal{I}^*_H} a(j_H, j^*_{H^c}) \frac{p(j_H, j^*_{H^c})}{p_\emptyset} \Big) \\ &= \Big(p_\emptyset \prod_{D \in \mathcal{D}} \prod_{i_D \in \mathcal{I}^*_D} p(i_D, i^*_{D^c}) \Big) \Big(1 - \frac{1}{p_\emptyset} \sum_{H \in \mathcal{U}} \sum_{j_H \in \mathcal{I}^*_H} a(j_H, j^*_{H^c}) p(j_H, j^*_{H^c}) \Big) \end{aligned} \tag{4.18}$$

The proof of this lemma is given in the Appendix.

We can now give the conjugate prior (3.1) in terms of p as given in (2.14). Let us note first that by (2.2) and (2.16), the marginal cell counts $y = (n(i_D), i_D \in \mathcal{I}^*_D, \; D \in \mathcal{D})$ for the multinomial distribution Markov with respect to G has density

$$\begin{aligned} f((y \mid p) &\propto \prod_{D \in \mathcal{D}} \prod_{i_D \in \mathcal{I}^*_D} \Big(\prod_{F \subseteq D} p(i_F, i^*_{F^c})^{(-1)^{|D \setminus F|}} \Big)^{y(i_D)} p_\emptyset^n \\ &\propto \prod_{D \in \mathcal{D}} \prod_{i_D \in \mathcal{I}^*_D} p(i_D, i^*_{D^c})^{u(i_D)} p_\emptyset^{u(\emptyset)} \end{aligned} \tag{4.19}$$

where $u(i_D) = \sum_{F \supseteq D} (-1)^{F \setminus D} \sum_{j_F \mid (j_F)_D = i_D} y(j_F)$ and $u(\emptyset) = n + \sum_{i_D \in \mathcal{I}^*_D, D \in \mathcal{D}} (-1)^{|D|} y(i_D)$.

Theorem 4.1 *For $(s, \alpha) \in \Pi_G$ and a_H as given in (4.16), the conjugate prior distribution induced from (3.1) by (4.17), that is, the conjugate prior for the parameter p of the multinomial family of distributions (4.19) is*

$$\pi^p_G(p \mid (s, \alpha)) \;\; = \;\; \frac{K^{-1}}{I_G(s, \alpha)} \prod_{D \in \mathcal{D}} \prod_{i_D \in \mathcal{I}^*_D} p(i_D, i^*_{D^c})^{\alpha(i_D, i^*_{D^c}) - 1} p_\emptyset^{\alpha_\emptyset - 1} \; . \tag{4.20}$$

where

$$\begin{aligned} K &= \Big(1 - \frac{1}{p_\emptyset} \sum_{H \in \mathcal{U}} \sum_{j_H \in \mathcal{I}^*_H} a(j_H, j^*_{H^c}) p(j_H, j^*_{H^c}) \Big) \\ \alpha(i_D, i^*_{D^c}) &= \sum_{F \supseteq D} \sum_{j_F \in \mathcal{I}^*_F \mid (j_F)_D = i_D} (-1)^{F \setminus D} s(i_F) \qquad (4.21) \\ \alpha_\emptyset &= \alpha + \sum_{i_D \in \mathcal{I}^*_D, D \in \mathcal{D}} (-1)^{|D|} s(i_D) \; . \qquad (4.22) \end{aligned}$$

This result follows immediately from the expression of the conjugate prior (3.1) in terms of θ, (2.2) and (4.18).

Example
When the graph is the four cycle with binary data as considered before

$$\mathcal{U}_0 = \{ac, bd, abcd, \emptyset\}.$$

From (4.14) and the constraints $\theta_E(i_E) = 0$ for $E \notin \mathcal{D}$, we have

$$a_{ac} = a_{bd} = 1, \;\; a_{abcd} = -1, \;\; \frac{p_{ac}}{p_\emptyset} = \frac{p_a p_c}{p_\emptyset^2}, \;\; \frac{p_{bd}}{p_\emptyset} = \frac{p_b p_d}{p_\emptyset^2}, \;\; \frac{p_{abcd}}{p_\emptyset} = \frac{p_{ab} p_{bc} p_{cd} p_{da}}{p_a p_b p_c p_d}$$

and

$$\begin{aligned}\pi(p_D, D \in \mathcal{D} \mid (s, \alpha)) = \\ I_G(s,\alpha)^{-1} p_a^{s_a - s_{da} - s_{ab} - 1} p_b^{s_b - s_{ab} - s_{bc} - 1} p_c^{s_c - s_{bc} - s_{cd} - 1} p_d^{s_d - s_{cd} - s_{da} - 1} p_{ab}^{s_{ab} - 1} p_{bc}^{s_{bc} - 1} p_{cd}^{s_{cd} - 1} p_{da}^{s_{da} - 1} p_\emptyset^{\alpha - 1} \\ (1 - \frac{p_a p_c}{p_\emptyset^2} - \frac{p_b p_d}{p_\emptyset^2} + \frac{p_{ab} p_{bc} p_{cd} p_{da}}{p_a p_b p_c p_d})^{-1}\end{aligned}$$

4.3 The conjugate prior for a general hierarchical model

When the model is not specified to be graphical but is a hierarchical log-linear model, we can also obtain the induced prior in terms of p and the statement is similar to Theorem 4.1 above except that the term coming from the Jacobian $|\frac{d\theta}{dp}|$ is more general and we have

Theorem 4.2 *For $(s, \alpha) \in \Pi_G$ the conjugate prior distribution induced from (3.1) by (4.17), that is, the conjugate prior for the parameter p of the multinomial family of distributions (4.19) for the hierarchical log-linear model is as in (4.20) with*

$$K \;\; = \;\; 1 - \sum_{H \in \mathcal{E}} \sum_{j_H \in \mathcal{I}_H^*} p(j_H, j_{H^c}^*) (\sum_{\{D \subseteq H, D \in \mathcal{D}\}} \sum_{\{C \subseteq D, C \in \mathcal{D}\}} (-1)^{|D \setminus C|})$$

and $\alpha(i_D, i_{D^c}^*)$ and $\alpha_\emptyset$ as in (4.21) and (4.22)

The proof follows immediately from the expression of the conjugate prior (3.1) in terms of θ, (2.2), (4.18) and Remark (8.1).

5 Properties of the conjugate prior

5.1 Hyper-parameter specification

Let us now turn to the practical problem of choosing hyperparameters which will reflect either some prior belief or lack of prior belief.

Suppose first that we do not have any prior information and want to put a flat prior on the log-linear parameters. From the expression (2.17) of the distribution of the marginal counts $y = (n(i_D), D \in \mathcal{D}, i_D \in \mathcal{I}_D^*)$, it is clear that the hyperparameters $s(i_D)$ can be thought of as the "prior marginal counts" for the marginal cell i_D. Therefore, we can take for $s(i_D)$ the set of i_D-"marginal counts" , $D \in \mathcal{D}, i_D \in \mathcal{I}_D^*$ for a "prior" contingency table with all "cell counts"' equal to $\frac{1}{\mathcal{I}}$. We can also take α to be the "total count", that is 1 in this case. This would lead, of course to

$$s(i_D) = \sum_{j \in \mathcal{I}, j_D = i_D \in \mathcal{I}_D^*} \frac{1}{\mathcal{I}} = \frac{1}{\mathcal{I}} \prod_{\gamma \in D^c} |\mathcal{I}_\gamma| \,,$$

where D^c is the complement of D in V. Since for the saturated model, the conjugate prior in terms of cell probabilities is the Dirichlet, it is clear that this choice of hyperparameters also yields a flat prior for the cell probabilities of the saturated model with all hyperparameters being equal to $\frac{1}{I}$. This prior is in fact the vague prior advocated by Perks (1947) (see also Dellaportas and Forster, 1999).

If we have prior information, we can first exclude all the interactions that are thought to be absent. Indeed, by Lemma 2.3, if two variables are believed to be independent given the others, then all $\theta_E(i_E) = 0$ for $E \in \mathcal{E}$ containing these two variables. We may have additional information such as the knowledge of positive or negative interaction between one or more variables. This knowledge can be expressed by computing the expected value and variance for e^{θ_D} for appropriate $D \in \mathcal{D}$. To illustrate what we mean, let us consider the data given by Hook, Albright and Cross (1980) and studied by King and Brooks (2001). In this data set, there are three variables

$$a \equiv BC, \;\; b \equiv DC, \;\; c \equiv MR$$

each taking the values 1 or 0 representing the presence or absence of, respectively, birth certificates, death certificates and medical records for each individual. The individuals under study are children with spina bifida. The data consists of an incomplete contingency table for each one of six years. From Hook, Albright and Cross (1980), it can reasonably be assumed that the model is the decomposable graphical model with cliques a and bc. Since the data is binary, from (2.22), the conjugate prior will then be of the form

$$\begin{aligned} \pi_G(\theta|s,\alpha) &= I_G(s,\alpha)^{-1} \exp\Big(\theta_a s_a + \theta_b s_b + \theta_c s_c + \theta_{bc} s_{bc} \\ &\quad -\alpha \log(1 + e^{\theta_a} + e^{\theta_b} + e^{\theta_c} + e^{\theta_a+\theta_b} + e^{\theta_a+\theta_c} + e^{\theta_b+\theta_c+\theta_{bc}} + e^{\theta_a+\theta_b+\theta_c+\theta_{bc}})\Big) \end{aligned} \tag{5.1}$$

There is also some prior knowledge about the interaction between b and c, that is for

$$e^{\theta_{bc}} = \frac{p_{bc}p_\emptyset}{p_b p_c}.$$

With high probability, $e^{\theta_{bc}}$ is expected to be in the interval $(-.9, -.1)$. From (5.1) and the formulas given in Proposition 4.1, if we let $s' = (s_a, s_b, s_c, s_{bc} + 1)$ we have

$$\begin{aligned} E(e^{\theta_{bc}}) &= \frac{I_G(s',\alpha)}{I_G(s,\alpha)} \\ &= \frac{\Gamma(s_{bc}+1)\Gamma(s_b - s_{bc} - 1)\Gamma(s_c - s_{bc} - 1)\Gamma(\alpha - s_b - s_c + s_{bc} + 1)}{\Gamma(s_{bc})\Gamma(s_b - s_{bc})\Gamma(s_c - s_{bc})\Gamma(\alpha - s_b - s_c + s_{bc})} \\ &= \frac{s_{bc}(\alpha - s_b - s_c + s_{bc})}{(s_b - s_{bc} - 1)(s_c - s_{bc} - 1)}\,. \end{aligned} \tag{5.2}$$

We therefore have the constraint

$$-.9 \le \frac{s_{bc}(\alpha - s_b - s_c + s_{bc})}{(s_b - s_{bc} - 1)(s_c - s_{bc} - 1)} \le -.1.$$

In the absence of any prior knowledge on the other log-linear parameters, we can assume that their expectation is around 0 which would imply that

$$\begin{aligned} E(e^{\theta_a}) &= \frac{\Gamma(\alpha - s_a - 1)\Gamma(s_a + 1)}{\Gamma(\alpha - s_a)\Gamma(s_a)} = \frac{s_a}{\alpha - s_a - 1} \\ E(e^{\theta_b}) &= \frac{\Gamma(\alpha - s_b - 1 - s_c + s_{bc})\Gamma(s_b + 1 - s_{bc})}{\Gamma(\alpha - s_b - s_c + s_{bc})\Gamma(s_b - s_{bc})} = \frac{s_b - s_{bc}}{\alpha - s_b - s_c + s_{bc} - 1} \\ E(e^{\theta_c}) &= \frac{\Gamma(\alpha - s_b - 1 - s_c + s_{bc})\Gamma(s_c + 1 - s_{bc})}{\Gamma(\alpha - s_b - s_c + s_{bc})\Gamma(s_c - s_{bc})} = \frac{s_c - s_{bc}}{\alpha - s_b - s_c + s_{bc} - 1} \end{aligned}$$

are all around 1. If we took all three ratios to be 1, we would obtain the relationships

$$2s_a = \alpha - 1,\ s_b - s_{bc} = s_c - s_{bc},\ 2(s_b - s_{bc}) = \alpha - s_c - 1,\ 2(s_c - s_{bc}) = \alpha - s_b - 1,$$
$$-.9 \leq \frac{s_{bc}}{s_b - s_{bc} - 1} \leq -.1\ . \tag{5.3}$$

and choose appropriate (s, α) satisfying these conditions. We might also want to compute the variance of these quantities which is, of course, also immediate with the results of Proposition 4.1, and give an interval where we wish $E(e^{\theta_D}), D \in \{a, b, c\}$ to be.

In general, when the model considered is not necessarily a decomposable graphical model, the ratio of normalising constants of the type $\frac{I_G(s',\alpha)}{I_G(s,\alpha)}$ has to be computed numerically. This is feasible by any or the standard MCMC or approximation methods. However, it might be wiser and much simpler to choose a decomposable model covering the interaction believed to be true. For example, if, in the example above, the prior model was believed to be the hierarchical model with generating class $\{ab, bc, ca\}$, then a reasonable prior model would be the saturated model Markov with respect to the complete graph subject to the fact that the interaction between a, b and c is weak, that is $E(e^{\theta_{abc}})$ is close to 1.

It remains to know whether the hyperparameters chosen for the conjugate prior on the log-linear parameter θ will yield hyper parameters in the conjugate prior induced by (3.1) for the cell probabilities which are consistent with the given prior beliefs. From Theorem 4.1, we know that the induced prior for the cell probabilities "'looks" like a Dirichlet on the free cell probabilities, that is $p(i_D, i^*_{D^c}), D \in \mathcal{D}, i \in \mathcal{I}^*_D$ with an additional factor for the Jacobian. The powers of the $p(i_D, i^*_{D^c})$ correspond to "prior cell counts" $n(i_D)$ and therefore any choice $s(i_D)$ in (3.1) will have the same meaning in (4.20). For example, corresponding to the condition that (5.2) be in the interval $(-.9, -.1)$ corresponds the condition that

$$E(\frac{p_{bc}p_\emptyset}{p_b p_c})$$

be in that interval also. From (4.20), the conjugate prior on $p = (p_a, p_b, p_c, p_{bc})$ is

$$\pi^p_G(p \mid (s, \alpha)) \quad = \quad \frac{\left(1 - \frac{p_a p_b + p_a p_c + p_a p_{bc}}{p^2_\emptyset}\right)^{-1}}{I_G(s, \alpha)} p_a^{s_a - 1} p_b^{s_b - s_{bc} - 1} p_c^{s_c - s_{bc} - 1} p_{bc}^{s_{bc} - 1} p_\emptyset^{\alpha - s_a - s_b - s_c + s_{bc} - 1}\ .$$

Therefore

$$E(\frac{p_{bc}p_\emptyset}{p_b p_c}) = \frac{I_G(s', \alpha)}{I_G(s, \alpha)}$$

where $s' = (s_a, s_b - 1, s_c - 1, s_{bc} + 1)$ and it follows immediately that

$$E(\frac{p_{bc}p_\emptyset}{p_b p_c}) = \frac{s_{bc}(\alpha - s_b - s_c + s_{bc})}{(s_b - s_{bc} - 1)(s_c - s_{bc} - 1)}\ ,$$

thus giving the same condition as in (5.2).

5.2 The strong hyper Markov property for local updates in graphical model

Let us now assume that the multinomial distribution of the contingency cell counts is Markov with respect to an arbitrary undirected graph G. We know from Dawid and Lauritzen (1993) that the multinomial distribution is strong meta Markov and as the conjugate distribution of the parameter θ of the exponential family (2.22), the conjugate prior (3.1) is strong hyper Markov.

Consider the decomposition of G into its prime components and let $P_l, l = 1, \ldots, k$ be a perfect enumeration of these components. Let $S_l, l = 2, \ldots, k$ be the corresponding separators. We now give the expression of (3.1) as the Markov ratio of conjugate priors on the prime components and the separators of G.

Proposition 5.1 *The conjugate prior (3.1) can be written as the Markov ratio*

$$\pi_G(\theta|s,\alpha) \quad = \quad \frac{\prod_{l=1}^k \pi_{P_l}(\theta^{P_l}|s^{P_l},\alpha)}{\prod_{l=2}^k \pi_{S_l}(\theta^{S_l}|s^{S_l},\alpha)} \tag{5.4}$$

where

$$\begin{aligned} &\pi_{P_l}(\theta^{P_l}|s^{P_l},\alpha) \qquad\qquad (5.5)\\ &= \quad I_G(s^{P_l},\alpha)^{-1}\exp\{\sum_{D\in\mathcal{D}_{P_l}}\sum_{i_D\in\mathcal{I}_D^*}\theta_D^{P_l}(i_D)s(i_D)\} - \alpha\log\Big(1+\sum_{i\in\mathcal{I}_{P_l},i\neq i^*}\exp\sum_{D\in\mathcal{D}_i^*}\theta_D^{P_l}(i_D)\Big) \end{aligned}$$

and where $s^{P_l} = (s(i_D),\ \mathcal{D}^{P_l}, i_D\in\mathcal{I}_D^*)$ *and* $s^{S_l} = (s(I_D),\ \mathcal{D}^{S_l}, i_D\in\mathcal{I}_D^*)$.
The induced conjugate prior (4.20) on p can be written as the corresponding Markov ratio of conjugate priors induced on p^{P_l} *and* p^{S_l} *from* $\pi_{P_l}(\theta^{P_l}|s^{P_l},\alpha)$ *in (5.5) and* $\pi_{S_l}(\theta^{S_l}|s^{S_l},\alpha)$.

Proof: It is not difficult to see that

$$\theta_D(i_D) = \sum_{l=1}^k \theta_D^{P_l}(i_D) - \sum_{l=2}^k \theta_D^{S_l}(i_D),\ D\in\mathcal{D} \tag{5.6}$$

where

$$\begin{aligned} \theta_D^{C_l}(i_D) \quad &= \quad \sum_{F\subseteq D, F\in\mathcal{D}_0^{C_l}} (-1)^{|D\setminus F|}\log p^{C_l}(i_F, i^*_{F^c}), \quad \text{for } D\subseteq C_l\\ \theta_\emptyset^{P_l} \quad &= \quad \log p_\emptyset^{P_l}\\ \theta_D^{P_l}(i_D) \quad &= \quad 0 \quad \text{for } D\not\subseteq P_l,\ l=1,\ldots,k, \qquad \theta_D^{S_l}(i_D)=0 \quad \text{for } D\not\subseteq S_l, l=2,\ldots,k, i_D\in\mathcal{I}_D^* \end{aligned}$$

We have proved this property for a decomposable graph G in §4.1, (4.10). The proof goes exactly along the same lines here. Therefore, if we let

$$\mathcal{D}^{P_l} = \{D\in\mathcal{D}\mid D\subseteq P_l\}, \quad \mathcal{D}^{S_l} = \{D\in\mathcal{D}\mid D\subseteq S_l\},$$

$$\theta^{P_l} = (\theta^{P_l}(i_D), D\in\mathcal{D}^{P_l}, i_D\in\mathcal{I}_D^*), \quad \theta^{S_l} = (\theta^{S_l}(i_D), D\in\mathcal{D}^{S_l}, i_D\in\mathcal{I}_D^*),$$

and

$$s^{P_l} = (s(i_D), D\in\mathcal{D}^{P_l}, i_D\in\mathcal{I}_D^*), \quad s^{S_l} = (s(i_D), D\in\mathcal{D}^{S_l}, i_D\in\mathcal{I}_D^*),$$

we see that

$$\sum_{D\in\mathcal{D}, i_D\in\mathcal{I}_D^*} s(i_D)\theta_D(i_D) \quad = \quad \sum_{l=1}^k \sum_{D\in\mathcal{D}^{P_l}, i_D\in\mathcal{I}_D^*}\theta_D^{P_l}(i_D)s(i_D) - \sum_{l=2}^k \sum_{D\in\mathcal{D}^{S_l}, i_D\in\mathcal{I}_D^*}\theta_D^{S_l}(i_D)s(i_D)\ .$$

Since by the Markov property, we also have $p_\emptyset = \frac{\prod_{l=1}^k p_\emptyset^{P_l}}{\prod_{l=2}^k p_\emptyset^{S_l}}$, that is

$$\begin{aligned} \log\Big(1+\sum_{i\in\mathcal{I},i\neq i^*}\exp\sum_{D\in\mathcal{D}_i^*}\theta_D(i_D)\Big) \quad &= \quad \sum_{l=1}^k \log\Big(1+\sum_{i\in\mathcal{I}_{P_l},i\neq i^*}\exp\sum_{D\in\mathcal{D}_i^*}\theta_D^{P_l}(i_D)\Big)\\ &\qquad -\sum_{l=2}^k\log\Big(1+\sum_{i\in\mathcal{I}_{S_l},i\neq i^*}\exp\sum_{D\in\mathcal{D}_i^*}\theta_D^{S_l}(i_D)\Big) \end{aligned}$$

it follows immediately that (5.4) is verified.

We note that as the restriction of s to $\mathcal{D}^{P_l}, i_D \in \mathcal{I}_D^*$, the coefficients of $(\theta^{P_l}(i_D), D \in \mathcal{D}^{P_l}, i_D \in \mathcal{I}_D^*)$ in (5.5) are consistent across prime components and separators. The factorization of the induced conjugate prior on p can be proved in a similar fashion. □

For given data y with total count n, the posterior distribution of θ given y will be

$$\pi_G(\theta|s+y,\alpha+n) \quad = \quad \frac{\prod_{l=1}^k \pi_{P_l}(\theta^{P_l}|s^{P_l}+y^{S_l},\alpha+n)}{\prod_{l=2}^k \pi_{S_l}(\theta^{S_l}|s^{S_l}+y^{S_l},\alpha+n)} \tag{5.7}$$

When comparing two models G an dG' the Bayes factor is the ratio of quantities of the type

$$\frac{I_{G'}(s,\alpha)}{I_{G'}(s,\alpha)}\,.$$

In the restricted class of decomposable graphical models, it is well-known that one can go from one decomposable graph to another through a succession of graphs that differ by only one edge. The additional edge can only belong to one clique in the new graph and as a consequence the Bayes factor affects only the graph induced by two cliques (see (37) in Dawid and Lauritzen, 1993). We are not aware of any such rule in the case of nondecomposable models. However, it is clear that the Bayes ratio will only involve the computation of the normalising constants for the subgraph induced by the prime components P_l affected by the additional edge.

6 Conclusion

In this paper we have studied the conjugate prior for the log-linear parameters of discrete hierarchical log-linear models and its induced prior on the cell probability parameter p thus extending the hyper Dirichlet which was the only form of the conjugate prior identified so far.

This prior has all the properties that one usually requires. As we have shown it, it has a moderate number of hyper-parameters precisely as many as there are log-linear parameters plus one. These hyperparameters are consistent across models. It is not difficult to translate prior knowledge into constraints for the hyper-parameters and constraints both in terms of the log-linear parameters and cell probabilities are consistent with prior beliefs, as illustrated in §5.1.

This prior has the additional property of being strong hyper Markov, thus leading to local updates for the computation of Bayes factors and it is also, of course, mathematically convenient since the prior and the posterior have the same form as the likelihood. The conjugate prior should therefore be one of the priors used for the study of contingency tables with a multinomial distribution for the cell counts. Though we have not mentioned it above, the translation of our results to the case of Poisson sampling is immediate.

7 References

Agresti, A. (1990). Categorical Data Analysis, *Wiley.*

Clyde, M. and George, E.I. (2004), Model uncertainty, *Statistical Science*, **19**, 81-94.

Consonni, G. and Leucari, V. (2006). Reference priors for discrete graphical models, *Bka*, **93**, 23-40.

Darroch, J.N., Lauritzen, S.L. and Speed, T.P. (1980). Markov Fields and log-linear models for contingency tables, *Ann. Statist.*, **8**, 522-539.

Dellaportas, P. and Forster, J.J. (1999). Markov chain Monte Carlo model determination for hierarchical and graphical log-linear models, *Bka*, **86**, 615-633.

Diaconis, P. & Ylvisaker, D. (1979). Conjugate priors for exponential families. *Ann. Statist.*, **7**, 269-81.

Hook, Albright and Cross (1980). Use of Bernoulli census and log-linear methods for estimating the prevalence of spina bifida in live births and the completeness of vital records reports in New York State. *Amer. J. Epidemiology*, **112**, 750-758.

King, R. and Brooks, S.P., (2001). Prior induction for log-linear models for general contingency table analysis, *Ann. Stat.*, **29**, 715-747.

Lauritzen, S. L. (1996). *Graphical Models*. Oxford University Press.

Madigan, D. and Raftery, A.E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *J.A.S.A.*, **89**, 1535-1546.

Perks, W. (1947). Some observations on inverse probability including a new indifference rule. *J. Insti. Actuar.*. **73**, 285-334.

8 Appendix

8.1 Proof of Lemma 4.1

We will first give the proof in the case where G is the simple decomposable graph $a--b--c$ and the data is bivariate. We will then sketch the proof for the general case of an arbitrary decomposable graph and discrete data.

In the particular case of bivariate data, p^G in Lemma 4.1 becomes

$$p^G = (p_D^{C_i}, D \in \mathcal{D}_{C_i} \setminus (\cup_{j=2}^k \mathcal{D}_{S_j}),\ i = 1, \ldots, k,\ \ p_D^{S_i}, D \in \mathcal{D}_{S_i},\ i = 2, \ldots, k) \tag{8.1}$$

and the Jacobian of the change of variables from $\theta = (\theta_D, D \in \mathcal{D})$ as given in (2.18) to p^G as given in (8.1) is

$$\Big|\frac{d\theta}{dp^G}\Big|^{-1} = \frac{\prod_{i=1}^k \prod_{D \in \mathcal{D}_0^{C_i}} p_D^{C_i}}{\prod_{i=2}^k \prod_{D \in \mathcal{D}_0^{S_i}} p_D^{S_i}} \tag{8.2}$$

We are therefore going to first prove (8.2) for the two-chain graph above. In this case we have $C_1 = \{a, b\}, C_2 = \{b, c\}, S = \{b\},\ \ p^G = (p_a^{C_1}, p_{ab}^{C_1}, p_c^{C_2}, p_{bc}^{C_2}, p_b^S)$ and

$$e^{\theta_a} = \frac{p_a}{p_\emptyset},\ e^{\theta_b} = \frac{p_b}{p_\emptyset},\ e^{\theta_c} = \frac{p_c}{p_\emptyset},\ e^{\theta_a + \theta_{ab}} = \frac{p_{ab}}{p_b},\ e^{\theta_{bc} + \theta_c} = \frac{p_{bc}}{p_b},$$

Moreover, since the multinomial distribution is Markov with respect to the graph G, we have

$$p_{abc} = \frac{p_{ab}p_{bc}}{p_b} \quad \text{and} \quad p_{ac} = \frac{p_a p_c}{p_\emptyset}.$$

Therefore

$$\frac{p_a^{C_1}}{p_\emptyset^{C_1}} = \frac{p_a + p_{ac}}{p_\emptyset + p_c} = \frac{p_a}{p_\emptyset}\frac{(1 + \frac{p_c}{p_\emptyset})}{(1 + \frac{p_c}{p_\emptyset})} = \frac{p_a}{p_\emptyset} \tag{8.3}$$

$$\frac{p_c^{C_2}}{p_\emptyset^{C_2}} = \frac{p_c}{p_\emptyset} \tag{8.4}$$

$$\frac{p_{ab}^{C_1}}{p_b^{C_1}} = \frac{p_{ab} + p_{abc}}{p_b + p_{bc}} = \frac{p_{ab}}{p_b}\frac{(1 + \frac{p_{bc}}{p_b})}{(1 + \frac{p_{bc}}{p_b})} = \frac{p_{ab}}{p_b} \tag{8.5}$$

$$\frac{p_{bc}^{C_2}}{p_b^{C_2}} = \frac{p_{bc} + p_{abc}}{p_b + p_{ab}} = \frac{p_{bc}}{p_b}\frac{(1 + \frac{p_{ab}}{p_b})}{(1 + \frac{p_{ab}}{p_b})} = \frac{p_{bc}}{p_b} \tag{8.6}$$

$$\frac{p_b^S}{p_\emptyset^S} = \frac{p_b + p_{ab} + p_{bc} + \frac{p_{ab}p_{bc}}{p_b}}{p_\emptyset + p_a + p_c + \frac{p_a p_c}{p_\emptyset}} = \frac{p_b}{p_\emptyset}\frac{(1 + \frac{p_{ab}}{p_b})(1 + \frac{p_{bc}}{p_b})}{(1 + \frac{p_a}{p_\emptyset})(1 + \frac{p_c}{p_\emptyset})} \tag{8.7}$$

We introduce the intermediate variables

$$v_a = \frac{p_a^{C_1}}{p_\emptyset^{C_1}},\ \ v_b = \frac{p_b^S}{p_\emptyset^S},\ \ v_c = \frac{p_c^{C_2}}{p_\emptyset^{C_2}},\ \ v_{ab} = \frac{p_{ab}^{C_1}}{p_b^{C_1}},\ \ v_{bc} = \frac{p_{bc}^{C_2}}{p_b^{C_2}}.$$

From (8.3) to (8.7), we have

$$v_a = e^{\theta_a},\ v_c = e^{\theta_c},\ v_{ab} = e^{\theta_a+\theta_{ab}},\ v_{bc} = e^{\theta_{bc}+\theta_c} \text{ and } e^{\theta_b} = v_b\frac{(1+v_a)(1+v_c)}{(1+v_{ab})(1+v_{bc})}$$

It is then immediate to see that

$$\Big|\frac{d\theta}{dv}\Big| = \prod_{D\in\mathcal{D}} v_D^{-1} = \frac{p_\emptyset^{C_1}p_\emptyset^S p_\emptyset^{C_2}p_b^{C_1}p_b^{C_2}}{p_a^{C_1}p_{ab}^{C_1}p_b^S p_c^{C_2}p_{bc}^{C_2}} \tag{8.8}$$

Moreover, since $p_b^S = p_b^{C_1} + p_{ab}^{C_1} = p_b^{C_2} + p_{bc}^{C_2}$, then $p_\emptyset^{C_1} = 1 - p_a^{C_1} - p_b^{C_1} - p_{ab}^{C_1} = 1 - p_a^{C_1} - p_b^S$ and similarly $p_\emptyset^{C_2} = 1 - p_c^{C_2} - p_b^S$. Therefore

$$v_a = \frac{p_a^{C_1}}{1-p_a^{C_1}-p_b^S},\ v_c = \frac{p_c^{C_2}}{1-p_c^{C_2}-p_b^S},\ v_{ab} = \frac{p_{ab}^{C_1}}{p_b^S - p_{ab}^{C_1}},\ v_{bc} = \frac{p_{bc}^{C_2}}{p_b^S - p_{bc}^{C_2}},\ v_b = \frac{p_b^S}{1-p_b^S} \tag{8.9}$$

and the matrix of the Jacobian $\left|\frac{dv}{dp^G}\right|$ is

$$\frac{dv}{dp^G} = \begin{pmatrix} \frac{1-p_b^S}{(1-p_a^{C_1}-p_b^S)^2} & 0 & 0 & 0 & 0 \\ 0 & \frac{p_b^S}{(p_b^{C_1})^2} & 0 & 0 & 0 \\ * & * & \frac{1}{(1-p_b^S)^2} & * & * \\ 0 & 0 & 0 & \frac{p_b^S}{(p_b^{C_2})^2} & 0 \\ 0 & 0 & 0 & 0 & \frac{1-p_b^S}{(1-p_c^{C_2}-p_b^S)^2} \end{pmatrix} \tag{8.10}$$

The Jacobian is equal to the product of the diagonal elements and since $1 - p_b^S = p_\emptyset^S$

$$\Big|\frac{dv}{dp^G}\Big| = \frac{(p_\emptyset^S)^2(p_b^S)^2}{(p_\emptyset^{C_1})^2(p_b^{C_1})^2(p_\emptyset^S)^2(p_b^{C_2})^2(p_\emptyset^{C_2})^2} = \frac{(p_b^S)^2}{(p_\emptyset^{C_1})^2(p_b^{C_1})^2(p_b^{C_2})^2(p_\emptyset^{C_2})^2} \tag{8.11}$$

Therefore

$$\Big|\frac{d\theta}{dp_G}\Big| = \frac{p_\emptyset^S p_b^S}{p_\emptyset^{C_1}p_a^{C_1}p_{ab}^{C_1}p_b^{C_1}p_\emptyset^{C_2}p_b^{C_2}p_{bc}^{C_2}p_c^{C_2}}\ , \tag{8.12}$$

which proves the lemma for the simple two-link chain graph considered.
For a general decomposable graph with bivariate data, if we write $S = \cup_{i=2}^k S_i$, then, the intermediate variables will be

$$v = (\frac{p_D^{C_i}}{p_{D\cap S}^{C_i}}, D\in\mathcal{D}_{C_i}\setminus(\cup_{j=2}^k\mathcal{D}_{S_j}),\ i=1,\ldots,k,\ \ \frac{p_D^{S_i}}{p_\emptyset^{S_i}}, D\in\mathcal{D}_{S_i},\ i=2,\ldots,k)$$

and the proof will follow the same lines as above.
In the case of discrete data, the proof follows the same line as the proof above with the following substitutions. For $D\in\mathcal{D}$,

θ_D becomes $(\theta_D(i_D), i_D\in\mathcal{I}_D^*)$
p_D; becomes $(p(i_D), i_D\in\mathcal{I}_D^*)$
$p_D^{C_l}$ and $p_D^{S_l}$ become $p^{C_i}(i_D)$ and $p^{S_l}(i_D)$ respectively, $i_D\in\mathcal{I}_D^*$
v_D becomes $(v_D(i_D), i_D\in\mathcal{I}_D^*)$
$p_\emptyset^{C_l} = 1 - \sum_{D\in\mathcal{D}} p_D^{C_l}$; becomes $p_\emptyset^{C_l} = 1 - \sum_{D\in\mathcal{D}}\sum_{i_D\in\mathcal{I}_D^*} p^{C_l}(i_D)$ and similarly for $p_D^{S_l}$

8.2 Proof of Lemma 4.2

For ease of notation, we will give the proof of the lemma in the case of binary data. Since for each $C \in \mathcal{D}$ and $H \in \mathcal{E}$ there is only one cell in $\mathcal{I}_C^*$ and $\mathcal{I}_H^*$ respectively, we will adopt the notation

$$F_{C,H} = F(i_C, j_H), \;\; C \in \mathcal{D}, H \in \mathcal{E} \; .$$

Let us first prove that if H is decomposable, then (4.15) is true. We proceed by induction on the number k of cliques of H. Let $\mathcal{C} = \{C_1, \ldots, C_k\}$ be a perfect ordering of the cliques of H.
If H is complete, that is $k = 1$, we consider two cases, the case where $|H|$ is even and the case where it is odd. For $|H| = 2p, p \in \mathbf{N}$, there are $n_e = \sum_{k=1}^{p} \binom{|H|}{2k}$ nonempty subsets of H of even cardinality and $n_o = \sum_{k=0}^{p-1} \binom{|H|}{2k+1}$ subsets of odd cardinality. Therefore

$$\sum_{C \subseteq_G H} (-1)^{|C|-1} = \sum_{k=1}^{2p} \binom{2p}{k} (-1)^{k+1} = (1-1)^{2p} - \binom{2p}{0} (-1)^1 = 1$$

and (4.15) is verified. We omit the proof for the case $|H| = 2p-1$ which is parallel to that of the previous case. Therefore (4.15) is verified for $k = 1$.
Let us now assume that H is decomposable but not complete, that is $k > 1$ and let us assume that (4.15) is true for any decomposable subset with $k-1$ cliques. It is well-know from the theory of decomposable graphs that, if we write $H_{k-1} = \cup_{j=1}^{k-1} C_j$, then $H = H_{k-1} \cup (C_k \setminus S_k)$ where $S_k = H_{k-1} \cap C_k$ is the k-th minimal separator in H. Therefore we have

$$\sum_{C \subseteq_G H} F_{C,H} = \sum_{C \subseteq_G H_{k-1}} F_{C,H} + \Big(\sum_{C \subseteq_G C_k} F_{C,H} - \sum_{C \subseteq_G S_k} F_{C,H} \Big) \; . \tag{8.13}$$

The first term on the right hand side of (8.13) is equal to 1 by our induction assumption while each one of the two other terms is also equal to 1 because both C_k and S_k are complete and therefore (4.15) is also verified for decomposable H.
Let us now prove that if H is not decomposable and connected, $\sum_{C \subseteq_G H} F_{C,H}$ cannot be equal to 1. If H is not connected and its connected components $H^{(1)}, \ldots, H^{(l)}$, for some $l \geq 2$, are all decomposable, we clearly have

$$\sum_{C \subseteq_G H} F_{C,H} = \sum_{j=1}^{l} \Big(\sum_{C \subseteq_G H^{(j)}} F_{C,H^{(j)}} \Big) \neq 1.$$

If H is not connected and its components are not all decomposable, this implies that there is a nondecomposable subset F_1 of G which can be separated from another subset F_2 of G but this contradicts our assumption that G is a prime component of G. So, this case does not occur.
If H is not decomposable and connected, consider its set of cliques $\{C_1, \ldots, C_k\}$. Since H is not decomposable, there is no perfect ordering of the cliques and therefore for any given ordering, there exist a nonempty subset $\mathcal{Q} \subseteq \{3, \ldots, k\}$ such that for $j \in \mathcal{Q}$, there is no $i < j$ in the given ordering of the cliques of H with $S_j = C_j \cap (\cup_{l=1}^{j-1} C_l) \subseteq C_i$ and therefore

$$S_j = C_j \cap (\cup_{l=1}^{j-1} C_l) = \oplus_{l=1}^{s_j} S_{j_l}, \quad 2 \leq s_j \leq j-1$$

where the S_{j_l} can be chosen to be disjoints, with $S_{j_l} \subseteq C_j \cap C_m$ for some $m \in \{1, \ldots, j-1\}$.

For $j \in \overline{\mathcal{Q}} = \{2, \ldots, k\} \setminus \mathcal{Q}$, there exists $i < j$ in the given ordering of the cliques of H such that $S_j \subseteq C_i$. Therefore

$$\sum_{C \subseteq_G H} F_{C,H} \;\; = \;\; \sum_{C \subseteq_G C_1} F_{C,H} + \sum_{j \in \overline{\mathcal{Q}}} \Big(\sum_{C \subseteq_G C_j} F_{C,H} - \sum_{C \subseteq_G S_j} F_{C,H} \Big) \tag{8.14}$$

$$+ \sum_{j \in \mathcal{Q}} \Big(\sum_{C \subseteq_G C_j} F_{C,H} - \sum_{l=1}^{s_j} \sum_{C \subseteq_G S_{j_l}} F_{C,H} \Big) \; . \tag{8.15}$$

The sums $\sum_{C\subseteq_G U} F_{C,H},\ U = C_1, C_j, S_j, j \in \overline{\mathcal{Q}}$ are all equal to 1 since each of $C_1, C_j, S_j, j \in \overline{\mathcal{Q}}$ are complete and connected and therefore the right hand side of (8.14) is equal to 1. For the same reason, on line (8.15), for $U = C_j, S_{j_l}, j \in \mathcal{Q}, l = 1, \ldots, s_j, \sum_{C\subseteq_G U} F_{C,H} = 1$. Since $s_j \geq 2$,

$$\sum_{C\subseteq_G C_j} F_{C,H} - \sum_{l=1}^{s_j} \sum_{C\subseteq_G S_{j_l}} F_{C,H} \leq -1,\ \ j \in \mathcal{Q}$$

and therefore the sum on line (8.15) is less than or equal to $-|\mathcal{Q}|$. It follows that

$$\sum_{C\subseteq_G H} F_{C,H} \leq 0$$

and in particular it cannot be equal to 1. The lemma is now proved.

8.3 Proof of Lemma 4.3

Here again, we will give the proof of the lemma for binary data and we will use the notation of §2.3. To shorten notation, we will write $E \subseteq_G F$ to indicate that $E \subseteq F$ and $E \in \mathcal{D}$.
It is more convenient to compute $|\frac{d\theta}{dp}|$, express it in function of θ and take its inverse. From the expression (2.21) of $p_D, D \in \mathcal{D}$, we have

$$\begin{aligned}
\frac{dp_D}{d\theta_D} &= \frac{e^{\sum_{E\subseteq_G D}\theta_E}}{1+\sum_{F\in\mathcal{E}} e^{\sum_{E\subseteq_G F}\theta_E}} - \frac{(e^{\sum_{E\subseteq_G D}\theta_E})\sum_{F\supseteq D, F\in\mathcal{E}} e^{\sum_{E\subseteq_G F}\theta_E}}{(1+\sum_{F\in\mathcal{E}} e^{\sum_{D\subseteq_G F}\theta_D})^2} \\
&= \frac{e^{\sum_{E\subseteq_G D}\theta_E}}{1+\sum_{F\in\mathcal{E}} e^{\sum_{E\subseteq_G F}\theta_E}}\Big(1 - \frac{\sum_{F\in\mathcal{E}, F\supseteq D} e^{\sum_{E\subseteq_G F}\theta_E}}{1+\sum_{F\in\mathcal{E}} e^{\sum_{D\subseteq_G F}\theta_D}}\Big) \\
&= p_D(1 - \sum_{F\in\mathcal{E}, F\supseteq D} p_F) \qquad (8.16) \\
\frac{dp_D}{d\theta_C} &= -\frac{(e^{\sum_{E\subseteq_G D}\theta_E})\sum_{F\in\mathcal{E}, F\supseteq C} e^{\sum_{E\subseteq_G F}\theta_E}}{(1+\sum_{F\in\mathcal{E}} e^{\sum_{D\subseteq_G F,}\theta_D})^2}, \qquad C \in \mathcal{D}, C \not\subset D \\
&= -p_D \sum_{F\in\mathcal{E}, F\supseteq C} p_F\ . \qquad (8.17) \\
\frac{dp_D}{d\theta_C} &= \frac{e^{\sum_{E\subseteq_G D}\theta_E}}{1+\sum_{F\in\mathcal{E}} e^{\sum_{E\subseteq_G F}\theta_E}} - \frac{(e^{\sum_{E\subseteq_G D}\theta_E})\sum_{F\in\mathcal{E}, F\supseteq C} e^{\sum_{E\subseteq_G F}\theta_E}}{(1+\sum_{F\in\mathcal{E}} e^{\sum_{D\subseteq_G F}\theta_D})^2}, \quad C \subset D, C \neq D, \\
&= p_D(1 - \sum_{F\in\mathcal{E}, F\supseteq C} p_F) \qquad (8.18)
\end{aligned}$$

We fix an arbitrary order of the elements of $\mathcal{D}$. From (8.16), (8.17) and (8.18), it follows that the matrix of the Jacobian is such that the column of partial derivatives of p_D is the vector with C-component

$$p_D\Big(\delta_{\{F\subseteq D\}}(C) - \sum_{F\in\mathcal{E}, F\supseteq C} p_F\Big),\ \ C \in \mathcal{D}\,,$$

where $\delta_{\{F\subseteq D\}}(C)$ is equal to 1 if $C \subseteq D$ and is equal to 0 otherwise. We note first that p_D is common to all components of the D column and therefore

$$J = \det A \prod_{D\in\mathcal{D}} p_D \qquad (8.19)$$

where A is the $|\mathcal{D}| \times |\mathcal{D}|$ matrix with entries

$$A_{C,D} = \delta_{\{F \subseteq D\}}(C) - \sum_{F \in \mathcal{E}, F \supseteq C} p_F, \quad C, D \in \mathcal{D}.$$

We note next that in the rows r_C corresponding to $C \in \mathcal{D}$ maximal with respect to inclusion, the C entry, on the matrix diagonal, is the only entry such that

$$\delta_{\{F \subseteq D\}}(C) \neq 0 \,,$$

and therefore, for C maximal, we can write

$$r_C = e_C - (\sum_{F \in \mathcal{E}, F \supseteq C} p_F)\mathbf{1}^t \,, \tag{8.20}$$

where e_C is the $\mathcal{D}$-dimensional row vector with components all equal to 0 except for the C component, and $\mathbf{1}^t$ is the $\mathcal{D}$-dimensional row vector with all its components equal to 1.
We finally note that if $C_1 \subset C_2$ for C_1, C_2 in $\mathcal{D}$, then

$$\{F \in \mathcal{E}, F \supseteq C_2\} \subset \{F \in \mathcal{E}, F \supseteq C_1\}$$

and therefore if, in the matrix A, for $C \in \mathcal{D}$ not maximal with respect to inclusion, we replace the row r_C by

$$\tilde{r}_C = r_C + \sum_{F \supset C, F \in \mathcal{D}} (-1)^{|F \setminus C|} r_F$$

we have

$$\tilde{r}_C = e_C - \Big(\sum_{F \supseteq C, F \in \mathcal{D}} (-1)^{|F \setminus C|} (\sum_{H \in \mathcal{E}, H \supseteq F} p_H) \Big) \mathbf{1}^t \,. \tag{8.21}$$

The determinant of A is clearly equal to the determinant of the matrix $\tilde{A}$ obtained from A by replacing r_C by $\tilde{r_C}$ whenever $C \in \mathcal{D}$ is not maximal with respect to inclusion. Using (8.20) and (8.21), we have

$$\tilde{A} = I_{|\mathcal{D}|} - U\mathbf{1}^t$$

where U is the column vector $U = (\sum_{F \supseteq C, F \in \mathcal{D}} (-1)^{|F \setminus C|} (\sum_{H \in \mathcal{E}, H \supseteq F} p_H),\ C \in \mathcal{D})$. It is well-known that

$$\det \tilde{A} = 1 - \mathbf{1}^t U$$

Therefore

$$\begin{aligned} \det \tilde{A} &= 1 - \sum_{C \in \mathcal{D}} \Big(\sum_{D \supseteq C, D \in \mathcal{D}} (-1)^{|D \setminus C|} (\sum_{H \in \mathcal{E}, H \supseteq D} p_H) \Big) \\ &= 1 - \sum_{H \in \mathcal{E}} p_H (\sum_{\{D \subseteq H, D \in \mathcal{D}\}} \sum_{\{C \subseteq D, C \in \mathcal{D}\}} (-1)^{|D \setminus C|}) \end{aligned} \tag{8.22}$$

According to (4.13), the coefficients of p_H in the expression above are the sum of the entries $F_{D,H} = \sum_{\{C \subseteq_G D\}} (-1)^{|D \setminus C|})$ in the column H of F. Moreover, by Lemma 4.2, this sum $\sum_{D \subset H, D \in \mathcal{D}} F_{D,H}$ is equal to 1 if and only if $H \in \mathcal{E}$ is decomposable, connected and nonempty. Since $1 = \sum_{F \in \mathcal{E}_0} p_F = \sum_{F \in \mathcal{U}_0} p_F + \sum_{F \notin \mathcal{U}_0} p_F$, we can write

$$\begin{aligned} \det \tilde{A} &= \sum_{F \in \mathcal{U}_0} p_F + \sum_{F \notin \mathcal{U}_0} p_F - \sum_{H \in \mathcal{E}} p_H (\sum_{\{D \subseteq H, D \in \mathcal{D}\}} F_{D,H}) \\ &= \sum_{F \in \mathcal{U}_0} p_F + \sum_{F \notin \mathcal{U}_0} p_F - \sum_{H \notin \mathcal{U}_0} p_H - \sum_{H \in \mathcal{U}_0} p_H (\sum_{\{D \subseteq H, D \in \mathcal{D}\}} F_{D,H}) \\ &= - \sum_{H \in \mathcal{U}_0} p_H \Big((\sum_{\{D \subseteq H, D \in \mathcal{D}\}} F_{D,H}) - 1 \Big) \\ &= - \sum_{H \in \mathcal{U}_0} a_H p_H \end{aligned} \tag{8.23}$$

From (8.19) and (8.23), we derive the first expression for J in (4.18). The other expressions are deduced by replacing the different p_F by their expression with respect to $(\theta_D, D \in \mathcal{D})$.

Remark 8.1 *When the model is not specified to be graphical but is more generally hierarchical, the proof above holds up to (8.22).*